



University of Groningen

Onder of boven de maat. Een methode voor het bepalen van de grens voldoende/onvoldoende bij studietoetsen

Wijnen, Wynandus Henricus Franciscus Wilhelmus

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:
1971

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Wijnen, W. H. F. W. (1971). Onder of boven de maat. Een methode voor het bepalen van de grens voldoende/onvoldoende bij studietoetsen. s.n.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

In dit onderzoek werd gezocht naar een methode voor het vaststellen van de grens tussen voldoende en onvoldoende bij studietoetsen. Allereerst werd het cesuurprobleem verduidelijkt met behulp van de begrippen klassifikatie, beslissing, waardering en vergelijking, waarna het onderzoeksgebied werd beperkt tot studietoetsen, zoals die worden afgenomen bij groepen bestaande uit ongeveer 30 personen of meer. (Hoofdstuk 1).

Beredeneerd werd, dat het cesuurprobleem wordt opgelost door impliciet of expliciet een drietal van elkaar te onderscheiden beslissingen te nemen:

- het kiezen van een referentiepunt, waarmee de feitelijke prestatie wordt vergeleken,
- het kiezen van een meeteenheid om de nog juist toelaatbare afstand tot het referentiepunt te definiëren,
- het kwantificeren van de nog juist toelaatbare afstand met behulp van de gekozen meeteenheid.

De al oude discussie over de noodzaak van ofwel absolute ofwel relatieve systemen bij het waarderen van studieprestaties werd beschreven. Na een vergelijking van de aangevoerde argumenten werd gekozen voor een relatief systeem. (Hoofdstuk 2).

Een viertal methoden voor de oplossing van het cesuurprobleem werden met behulp van de te onderscheiden beslissingen geanalyseerd: de voorstellen van resp. Nedelsky, De Groot, "grading on the curve", en Ebel. Bij elk van deze methoden bleven enkele vraagtekens staan, waardoor een aanvaardbare oplossing door geen van de genoemde methoden kon worden bereikt. Het absolute referentiepunt bij Nedelsky en De Groot, en de enigszins willekeurige kwantificering van de toelaatbare afstand bij de relatieve methoden van "grading on the curve" en Ebel waren de belangrijkste bezwaren. Verder bleek dat in alle gevallen de kwaliteit van de verzamelde informatie (betrouwbaarheid) buiten beschouwing werd gelaten. (Hoofdstuk 3).

Voor de bezwaren, die tegen de verschillende oplossingsmethoden werden ingebracht, kon enige steun worden gevonden bij studietoetsen, zoals die momenteel in het wetenschappelijke onderwijs gebruikelijk zijn. (Hoofdstuk 4).

Uitgaande van de overtuiging, dat onderwijs zich behoort aan te passen aan degenen, die tot het volgen van dat onderwijs gekwalificeerd zijn, werd vervolgens een methode voor het bepalen van de grens tussen voldoende en onvoldoende voorgesteld, die bestaat uit de volgende elementen:

- het gemiddelde van de groep is het referentiepunt,
- de nog juist toelaatbare afstand wordt - rekening houdend met de meetkwaliteit - gedefinieerd in standaardmeetfouten,
- de kwantificering van de nog juist toelaatbare afstand - twee standaardmeetfouten - sluit aan bij het gebruikelijke 95%-betrouwbaarheidsinterval.

De beslissingsregel wordt dan, dat iedereen wordt afgewezen, van wie de feitelijke prestatie lager is dan het gemiddelde van de groep verminderd met twee standaardmeetfouten. (Hoofdstuk 5).

Het toepassen van de voorgestelde beslissingsregel op 45 studietoetsen wees uit, dat de beslissingsregel in de huidige studietoetspraktijk tot variatie in het percentage afwijzingen leidt, terwijl tevens bleek, dat het percentage afwijzingen niet zonder meer toeneemt, indien men de studietoets verlengt. Ook kon worden gekonstateerd, dat het afwijzingspercentage enigszins toeneemt, naarmate de studietoets moeilijker wordt. Een relatie tussen het afwijzingspercentage en de koëfficiënt voor interne consistentie kon eveneens worden aangetoond: naarmate de interne consistentie van de studietoets hoger wordt, neemt ook het afwijzingspercentage toe. (Hoofdstuk 6).

Wanneer een studietoets onverwacht weinig goede antwoorden oplevert, kunnen daaraan tenminste twee oorzaken ten grondslag liggen: de gestelde vragen zijn te moeilijk ofwel de inspanningen en/of capaciteiten van de kandidaten zijn te gering. Met behulp van een onderscheid tussen selektietoetsen en voortgangstoetsen, gebaseerd op een viertal ideaaltypische studietoetsen, werd een mogelijkheid tot differentiatie tussen beide toetsvormen beschreven. Dit leidde tot een operationalisatie van onderwijseffekten als variantie-bijdrage van de vraaggemiddelden en tot operationalisatie van persoonseffekten als variantie-bijdrage van de persoonsgemiddelden. (Hoofdstuk 7).

Resultaten uit "post-hoc"-experimenten gaven steun aan de gekozen operationalisaties. Bovendien kon worden vastgesteld, dat de studietoetsen, die momenteel op universitair niveau worden afgenomen, beter gekarakteriseerd kunnen worden als voortgangstoetsen dan als selektietoetsen, d.w.z. dat er meer informatie beschikbaar is over verschillen tussen de gestelde vragen dan informatie over verschillen tussen kandidaten. (Hoofdstuk 8).

Het selekteren c.q. elimineren van vragen uit studietoetsen kan eveneens op tenminste twee verschillende manieren gebeuren. Geargumenterd werd voor een selectieprocedure, die gebaseerd is op het extreem zijn van de p-waarde ten opzichte van de gemiddelde p-waarde over alle vragen. In dat geval wordt

het effect van itemselectie ten aanzien van de onderwijseffekten voorspelbaar, terwijl niet vast staat, wat er met het afwijzingspercentage gebeurt. Selekteert men op basis van de item-test-korrelatie, dan is het effect op het afwijzingspercentage voorspelbaar, terwijl niet zeker is, dat de onderwijseffekten hierdoor worden verkleind. Selectie gebaseerd op het extreem zijn van de p-waarde werd daarom beter geacht. (Hoofdstuk 9).

De in de literatuur beschreven relatie tussen standaardmeetfout en het aantal vragen in een studietoets kan niet zonder meer worden gehandhaafd. De relatie blijkt vooral dan te bestaan, indien studietoetsen een gemiddelde moeilijkheidsgraad van $p = .50$ naderen, terwijl de korrelaties tussen de vragen laag zijn. Beide voorwaarden zijn bij selectietoetsen minder wenselijk.

Wel moest worden geconstateerd, dat het bepalen van betrouwbaarheidsintervallen rond de feitelijke prestatie minder juist is dan het bepalen van betrouwbaarheidsintervallen ten opzichte van de betrouwbare score, behorend bij de feitelijke prestatie. Gezien de grotere praktische bruikbaarheid van een uniforme kwantificering van de toelaatbare afstand tot het referentiepunt, werd als konsekwentie hiervan niet de gekozen kwantificering herzien, maar de verbalisering van de beslissingsregel, die daardoor luidt: Iemand wordt afgewezen, wanneer met een waarschijnlijkheid van 95% kan worden gezegd, dat zijn prestatie niet kan worden gezien als een toevalsafwijking van een prestatie gelijk aan het groepsgemiddelde. (Hoofdstuk 10).

De wellicht voor de hand liggende redenering, dat de voor- gestelde beslissingsregel bij perfect betrouwbare studietoetsen ongeveer 50% van de kandidaten afwijst kon worden ontkracht door aan te tonen, dat perfecte betrouwbaarheden slechts mogelijk zijn bij zeer merkwaardige frekwentieverdelingen, terwijl bij meer gangbare verdelingen de betrouwbaarheid verre van perfect is. Tevens kon worden aangetoond, dat de stelling, dat de helft van de kandidaten een prestatie heeft lager dan het gemiddelde, in zijn algemeenheid niet juist is. (Hoofdstuk 11).

In een afsluitend hoofdstuk werd de beslissingsregel toegepast op een drietal studietoetsen. Hoewel deze toepassingen enkele problemen ten aanzien van itemselectie en variantie-bijdragen zichtbaar maakten, werden de principiële uitgangspunten van de in dit onderzoek gevolgde gedachtengang meer geïllustreerd dan tegengesproken. Tenslotte werden twee uitzonderingen genoemd, waarbij het toepassen van de beslissingsregel strijdig kan zijn met de onderwijsopvatting, die eraan ten grondslag ligt. (Hoofdstuk 12).